图书情报工作
第62卷 第3期 2018年2月

基于LDA主题模型的标签混合推荐研究*

■ 熊回香　窦燕
华中师范大学信息管理学院　武汉 430079

摘要:［目的/意义］针对目前使用标签推荐方法所得结果不理想的问题,改进传统相似度计算方式,并结合多种标签推荐方法,提高推荐准确性。［方法/过程］融合基于内容与协同过滤的推荐思想,利用LDA进行相似度计算得出资源与用户的近邻集合,并抽取资源内容关键词,以此构建标签混合推荐模型,最后以“豆瓣读书”为例对模型进行验证,同时与几种标签推荐方法进行比较。［结果/结论］在社会标注系统中,必须考虑用户-资源-标签3个维度,仅考虑单一角度势必会造成结果的不完整,同时在相似度计算时引入LDA能够挖掘潜在语义关系,提高推荐质量,且组合多种方法取长补短可以令推荐结果更为满意。

关键词:社会标注　标签推荐　协同过滤　LDA
分类号:TP181
DOI:10.13266/j.issn.0252-3116.2018.03.013

1 引言

社会标注是Web2.0时代一种主要且有效的网络信息资源组织方式,它允许用户使用自定义的关键词即标签来注释网络中的各种资源,以便有效组织、检索和利用这些资源。标签由用户创造,不受任何限制,一方面它反映了用户对资源的认识,另一方面通过标签用户可以检索资源或是寻找相同兴趣的用户[1]。社会标注在互联网中得到广泛应用的同时也产生了许多问题。例如,由于用户的偏好差异,不同用户会使用不同的标签来标注同一资源;而社会标签的无控性使得用户在自由标注时会产生错误标签或是无实际意义的垃圾标签[2],诸多问题成为网络信息组织与检索的障碍,也在一定程度上降低了标签作用的有效性。而标签推荐作为一种有效的方法,能够在用户对资源进行标注时为其推荐标签,既能改善用户体验,也能对用户的标注行为产生规范约束,从而提高标签质量[3]。目前国内外关于社会标注的研究主要集中于资源推荐与用户推荐,而关于标签推荐的研究较少,且研究也仅限于如基于内容的标签推荐、基于协同过滤的标签推荐或是基于关联规则的标签推荐等单一技术,组合多种技术的研究不多,标签系统中的标签有多种来源,仅使用单一推荐方法会使结果片面;另外,这些传统推荐技术未考虑到标签间包含的丰富语义信息或是用户因各种因素产生的偏好差异,从而导致在推荐质量上仍有所欠缺。因此,本文提出一种组合多种技术的标签混合推荐方法,该方法将基于内容的标签推荐、基于用户的协同过滤与基于资源的协同过滤3种方法相结合,同时将隐含狄利克雷分布(Latent Dirichlet allocation, LDA)引入到相似度计算过程中,以主题概率分布作为计算依据来取代传统的计算方式,加入深层语义知识,产生基于相似资源和基于相似用户的推荐标签,其次抽取资源内容的关键词作为基于内容的推荐标签,最后融合3种结果为用户进行标签推荐。其意义在于从3种标签来源角度出发,融合多种推荐技术,使推荐方法与标签来源角度一一对应,提高数据稠密性,从而避免单一技术的缺陷。经实验发现,这种基于LDA的混合标签推荐方法一定程度上缓解了标签语义模糊性等问题,推荐结果也有较大改进。

2 标签推荐及相关技术

2.1 标签推荐

社会标签既是用户兴趣偏好的代表,又能够从不同维度揭示资源特征,但其过度自由的特性导致系统

* 本文系国家社会科学基金项目“大众分类中标签间语义关系挖掘研究”(项目编号:12BTQ038)研究成果之一。
作者简介:熊回香(ORCID:0000-0001-9956-3396),教授,博士,E-mail:hxxiong@mail.ccnu.edu.cn;窦燕(ORCID:0000-0003-0029-2624),硕士研究生。
收稿日期:2017-08-25　修回日期:2017-11-19　本文起止页码:104-113　本文责任编辑:刘远颖

中出现了许多低质量标签,在一定程度上影响了标签作用的发挥[4]。为了提高标签质量,最直接的方法就是控制用户对标签的使用。但这种方法因其强迫性、限制性,必然会难以被用户接受,同时也不符合社会标注的自由性。因此,标签推荐机制应运而生[5]。标签推荐是指当某一用户想要对某个资源进行标注时,系统结合用户的标注情况、资源内容特征以及系统中已有标签等信息,为其推荐一系列相关标签进行选择。标签推荐作为用户标注时的一种辅助工具,一方面可以为其提供参考和建议,减轻用户负担,提升用户标注的积极性,另一方面也可以提高标签的质量,提高资源检索的效率和准确率[6]。相比强制控制用户对标签的使用,利用标签推荐这种更为友好和温和的建议式交互语言来适当规范用户的标注行为,简化了标注过程,改善了用户体验,更提高了标注结果的质量,十分具有研究意义[7]。

2.2 个性化推荐技术

目前,社会标注系统中标签推荐方法实际上都借鉴了电子商务领域中的个性化推荐技术。国内现今的个性化推荐技术主要有3种:基于内容的推荐、协同过滤的推荐和混合推荐。基于内容的推荐其推荐依据来自于资源内容本身,通常使用的是文本内容[8]。协同过滤分为两种,一是基于资源的协同过滤,二是基于用户的协同过滤。这种方法主要是通过计算用户或是资源间的相似度从而进行推荐[9]。由于每种推荐技术各有优劣,因而混合推荐近年来被频繁提出,多种推荐方式的取长补短有利于规避单一方法的缺点,提高推荐质量。当然,随着技术的进步推荐系统如今又出现了一些新方法,如基于关联规则的推荐、基于链路预测的推荐、基于社会网络信任关系的推荐等。

现今,许多学者都在上述推荐技术基础上结合标签特性提出了各式各样的标签推荐技术。国外学者M. Tatu等[10]将近邻法与关键词提取法相融合进行基于词标签的推荐;G. Mishne[11]利用K近邻法从文档集中挑选出与待标注资源最相关的K个资源,并将其标签推荐给用户;L. Marinho等[12]利用标签-用户-资源三者的关系得到两两间的二维矩阵,从而通过矩阵向量发现相似用户并进行标签推荐;A. Hotho等[13]则根据PageRank算法提出了FolkRank算法,利用链接分析实现推荐。在国内,宋洪鑫等[14]则对新浪博客标签进行了相关研究,提出了基于关键词提取与博客文章分类的推荐模型;高兵[15]关注的是问答式社区的标签推荐技术,根据待标引问题的文本寻找相似且已标引的问题进行推荐;王传豹[16]提出了基于协同过滤和文本相似度的混合标签推荐方法;安志伟[17]提出了一种基于三部图张量分解法的标签推荐算法;张亮[18]融合用户、标签、资源三者间的关系,直接利用LDA构建统一主题模型进行标签推荐。

由此可见:①在标签推荐系统中推荐技术大多为基于内容、基于协同过滤或是基于其他一些方法,角度单一,混合多种方法的推荐较少;②协同过滤、张量分解等方法是从用户-资源-标签三者间的关系出发,这种仅分析对象间关系的技术忽略了标签的语义与资源内容特征;③标签推荐对于LDA的应用,主要集中于利用LDA对标签进行聚类、语义分析以及直接利用LDA进行推荐,极少将LDA与其他标签推荐方法相结合使用。因而,本文针对以上问题,从多角度出发,结合目前使用较多的推荐技术,将对象间关系与标签语义内容同时考虑,并借助LDA改进传统的相似度计算,为用户进行标签推荐。

2.3 LDA主题模型

LDA是一种无监督的概率主题模型,常被用来对大规模文档集合进行建模。其基本思想基于一个假设:一个用户在写一篇文档时,心中必定会有一些确定的主题,有了主题后用户就必定会从某个主题的所有单词池中以一定的概率选择一个词来阐释该主题,整个文档就相当于不同主题的混合[1]。LDA核心思想如公式(1)所示:

$$p(\text{词语} \mid \text{文档}) = \sum_{\text{主题}} (\text{词语} \mid \text{主题}) \times p(\text{主题} \mid \text{文档}) \qquad \text{公式(1)}$$

LDA实质上是一个以文档-主题-词汇为层次结构的三层贝叶斯概率模型,如图1所示:

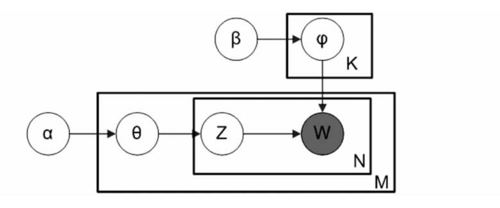


图1 LDA模型

其采用Dirichlet分布作为概率主题模型多项分布的先验分布,在该模型中,W代表词汇,是唯一可观测的变量[19],M代表整个文档集,N代表每篇文档包含的总词数,K代表主题个数,α和β分别代表文档-主题概率θ和主题-词语概率分布φ的先验分布超参数[20]。

LDA采用词袋的思想,先以一定的概率选取某个

主题，再以一定的概率选取该主题下的某个单词，不断重复以上步骤直至产生文档中所有的词语[21]。这种方法间接地对词汇进行模糊聚类，通过训练得到每篇文档在主题上的分布和每个主题在词空间上的分布，从而挖掘文本信息，既能衡量各文档间的潜在语义关系，又有强大的降维能力，缓解数据稀疏问题[22]。因此，在相似度计算时引入 LDA 主题模型，即便用户使用了不同的标签，或是资源由不同的特征词所表示，只要这些词属于相同主题，就可以很好地度量相似性，提高稀疏环境下的推荐质量[23]。

3 推荐框架描述与数据预处理

3.1 标签推荐模型描述

目前，标签推荐方法虽多种多样，但每种方法都有其无法避免的缺点，另外标签推荐系统与普通电子商务领域中的推荐系统不同，其含有三大元素，即用户－标签－资源，标签作为中介将用户与资源联系起来，因而也包含着资源内容标签、资源热门标签、用户兴趣标签3种标签来源，且其特点各不相同[24]，如果仅从某一个角度进行推荐，会导致数据有所缺失，必然会造成结果的偏差与不全面。因此，针对如何把多种标签来源融合在一起从而提升标签推荐的准确性，本文就以混合推荐为突破口提出了该推荐模型（见图2），其融合的原理就在于首先依据不同的标签来源采用不同的推荐技术，如从资源内容标签角度出发采用基于内容的推荐，从资源热门标签角度出发使用基于资源的协同过滤，而从用户兴趣标签角度看则使用基于用户的协同过滤，然后将这3种推荐技术的结果相融合，则最终产生的推荐结果必定包含所有标签来源，覆盖用户－标签－资源三方面，同时该方法结合对象间的关系与标签的语义进行分析，将数据粗粒度化使其变得更加稠密，从而避免了单一方法的不足。

总体推荐框架见图2，共5个阶段——数据收集、数据预处理、LDA训练、相似度计算和推荐结果生成。该模型中，假设用户集合U中有m名用户，图书资源集合R中有n本图书，标签集合T中有p个标签，针对用户 $u \in U$ 在标注资源 $r \in R$ 时，系统通过建立资源－词语语料库和用户－标签语料库，使用LDA得到资源和用户在主题上的分布概率从而进行相似度计算，并寻找用户u的近邻用户集和资源r的近邻资源集，最后将从相似资源与相似用户中得到的推荐标签与从资源中抽取的关键词相结合，为用户u推荐相关标签 $t \in T$。由于本文数据量很少，因此文中对所提到的k值等相关设定不做讨论，重点在于清晰论述标签推荐模型的构建及其实现，同时根据国内外相关文献，一般LDA模型的参数取值 $\alpha = 50/k$（k为主题个数），$\beta = 0.01$[25]，本文就以此参数为基准进行建模。

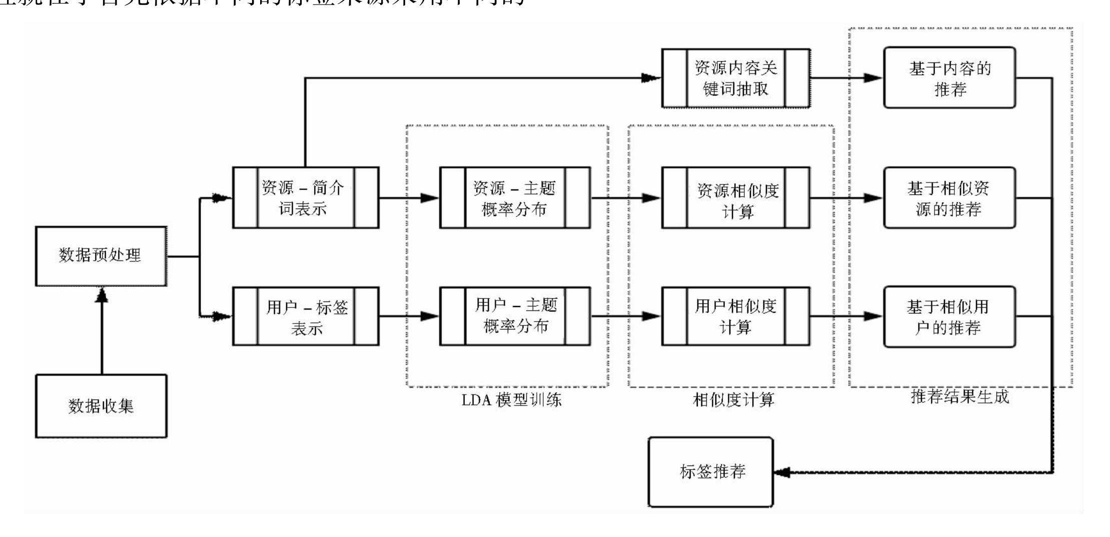


图2 标签推荐模型框架

3.2 实验数据

3.2.1 数据收集 豆瓣网是国内较为热门的社会标注网站，在该网站中，用户能为自己所喜欢的资源进行标注，标注范围主要包括图书、音乐、电影等方面，用户可以通过标签浏览标注了同一标签的资源找到新的兴趣点，也能通过标签找到与自己标注同一资源的用户，寻找兴趣相同的伙伴[26]。因此本文以豆瓣网的"豆瓣读书"频道为研究对象，结合相关分析过程阐述标签推荐模型。本文从"豆瓣读书"频道中通过人工浏览采集的方式随机选取了25名用户，并从这25名用户的

"在读""想读""读过"3个栏目中获取其图书名称、图书简介、每本图书的常用标签(10个)和每个用户标注相关图书的标签这些数据作为实验研究的基础。

3.2.2　数据预处理　首先利用中国科学院的NLPIR汉语分词系统对图书简介及不规范的标签进行分词，并利用停用词表过滤掉无实际意义的词(如"啊""就""嘿"等)以及一些特殊符号。同时，对于含有英文的标签，一律将其转换成小写形式。另外，图书名称、图书作者名等专有名词是描述资源特征的重要部分，也是标签推荐的关键来源，因而可以利用NLPIR的用户自定义词典功能，将这些词添加进自定义词典中，在分词处理时予以保留。经处理后，得到的数据为:25名用户、135本图书资源及其内容简介、592个图书常用标签和234个用户标签，如表1、表2所示:

表1　图书资源、简介及标签数据集[27]

编号	图书名称	简介内容	标签
1	《昨日的世界》	饱满;真挚;……;心迹	茨威格;传记;……;新知三联书店
2	《大数据经济》	中国;互联网;……;日常	大数据;互联网;……;创业
3	《毛泽东传记》	罗斯·特里尔;……;形象	传记;毛泽东;……人物
4	《总统总是靠不住》	信件;美国;……;监督	林达;美国;……;启蒙
5	《浩荡两千年》	中国;企业;……;企业	吴晓波;商业;……;经济史
6	《乡土中国》	社会学家;费孝通;……;研究者	社会学;费孝通;……;经典
7	《菊与刀》	本尼迪克特;菊与刀;……;文化	日本;文化;……;美国
8	《全球通史》	世界;历史;……;时期	历史;世界史;……;全球
9	《朱元璋传》	朱元璋;历史;……;基础	传记;历史;……;中国
10	《中国大历史》	中国;历史;……;衡量	历史;黄仁宇;……;历史学
11	《影响力》	政治家;影响力;……;就范	心理学;影响力;……;经济
12	《黄金时代》	文革;时期;……;阴影	王小波;黄金时代;……;大陆
13	《挪威的森林》	动人心弦;平缓;……;人生	村上春树;挪威的森林;……;文学
14	《深度案例思考法》	思考;逻辑;……;能力	思维;方法论;……;后浪
15	《硅谷之谜》	颠覆;信息;……;长盛不衰	互联网;吴军;……;文化
……	……	……	……
135	《1984》	1984;杰出;……;经典	乔治·奥威尔;反乌托邦;……;英国文学

表2　用户、图书及其标签数据集[27]

编号	用户	图书名称	标签
A	路过蜻蜓	《菊与刀》	日本;文化;社会;美国;历史
		《人类简史》	历史;文化;社会
		……	……
		《万历十五年》	历史;中国
B	Banyan	《乡土中国》	费孝通;乡土中国;社会学;人类学;经典
		《浪潮之巅》	互联网;IT;商业;历史
		……	……
		《挪威的森林》	村上春树;日本
C	飞扬	《人民的正义》	中国;周梅森;政治;小说
		《武则天传》	林语堂;中国;历史;小说
		……	……
		《天龙八部》	金庸;武侠;小说;中国
……	……	……	……
Y	binnng	《全球通史》	历史;世界史
		《从0到1:开启商业与未来的秘密》	美国;经济学;互联网
		……	……
		《长尾理论》	经济;经济学

4　基于LDA的标签混合推荐

4.1　资源主题模型训练及计算

4.1.1　资源－主题模型训练　经过3.2节数据预处理后，对于任意一本图书资源r都由n个简介词w表示，见表3。把图书看作文档，简介词看作文档中的词语，在此基础上利用LDA进行建模，可以得到资源－主题概率分布及主题－词语概率分布，而本文后续计算只需用到资源在主题上的概率分布。利用python及其LDA工具包对表3进行模型训练，取主题数k=15，经训练后得到资源－主题概率分布，即每一个资源中会出现该主题的概率大小，结果见表4。

表3　资源－词语矩阵

图书编号	简介特征词
1	饱满;真挚;感情;……;心迹
2	中国;互联网;行业;……;日常
3	罗斯·特里尔;叙述;……;形象
4	信件;美国;总统;……;监督
5	中国;企业;研究;……;企业
6	社会学家;费孝通;教授;……;研究者
7	本尼迪克特;菊与刀;日本;……;文化
……	……
135	1984;杰出;政治;……;经典

表 4 资源－主题概率分布矩阵

图书编号	topic1	topic2	topic3	topic4	……	topic15
1	0.003 39	0.071 19	0.376 27	0.003 39	……	0.071 19
2	0.026 75	0.396 18	0.001 27	0.064 97	……	0.001 27
3	0.001 98	0.001 98	0.239 60	0.021 78	……	0.001 98
4	0.004 26	0.004 26	0.004 26	0.046 81	……	0.089 36
5	0.140 87	0.210 43	0.036 52	0.071 30	……	0.001 74
6	0.002 74	0.002 74	0.030 14	0.002 74	……	0.002 74
7	0.001 34	0.001 34	0.014 77	0.001 34	……	0.001 34
……	……	……	……	……	……	……
135	0.002 99	0.002 99	0.002 99	0.480 60	……	0.331 34

4.1.2 相似度计算

（1）主题概率距离。传统的协同过滤技术中常使用余弦相似度或是皮尔逊相关系数来计算相似度，但由于本文的计算依据是资源在主题概率上的分布，因此不能直接使用传统的计算公式。KL（Kullback-Leibler）散度，又称 KL 距离，常用来计算两个概率分布之间的距离，如公式（2）所示：

$$D_{kl}(p,q)=\sum_{i=1}p_i\ln\frac{p_i}{q_i}\qquad \text{公式(2)}$$

公式（2）中 p 和 q 为两个概率分布，且对于任意 i，当 $p_i=qi$ 时，$D_{kl}(p,q)=0$。但 *KL* 散度是一个非对称性距离，因此，为了便于计算，往往使用其对称公式 *JS*（(*Jensen-Shannon*）散度，如公式（3）所示：

$$D_{js}(p,q)=\frac{1}{2}\left[D_{kl}\left(p,\frac{p+q}{2}\right)+D_{kl}\left(q,\frac{p+q}{2}\right)\right]\qquad \text{公式(3)}$$

公式（3）中 *p* 和 *q* 同样为两个概率分布，该式的区间为[0,1][23]，即 *JS* 散度值越趋向于 0，则两个概率间的距离越近，值越趋向于 1，则表明两个概率距离越远。

本文选择此距离公式，以表 4 为基础，计算出两本图书资源之间主题概率分布的距离，结果如表 5 所示：

表 5 资源间概率分布距离矩阵

图书编号	1	2	3	4	……	135
1	0	0.398 48	0.136 94	0.324 58	……	0.511 19
2	0.398 48	0	0.336 17	0.442 71	……	0.514 7
3	0.136 94	0.336 17	0	0.389 06	……	0.585 57
4	0.324 58	0.442 71	0.389 06	0	……	0.402 36
5	0.311 61	0.206 78	0.264 77	0.374 69	……	0.499 62
6	0.326 73	0.382 39	0.241 02	0.350 32	……	0.505 9
7	0.454 44	0.491 91	0.383 44	0.467 81	……	0.459 16
……	……	……	……	……	……	……
135	0.230 89	0.572 91	0.318 84	0.614 7	……	0

（2）矩阵转换。表 5 为图书资源间的距离差异，根据公式（3）的描述可知数值越小两个个体间距离越近，为方便后续计算，需将其利用公式（4）转换成相似度矩阵。

$$Sim(a,b)=\frac{1}{(1+D(a,b))}\qquad \text{公式(4)}$$

在公式（4）中，Sim（a,b）为图书资源 a 和 b 之间的相似度，D（a,b）为 a 和 b 之间的主题分布距离，分母加 1 是为了防止距离为 0 时带来的影响，Sim 值越大，表示二者越相似。通过公式（4）计算结果如表 6 所示：

表 6 资源相似矩阵

图书编号	1	2	3	4	……	135
1	1	0.715 06	0.879 55	0.754 96	……	0.661 73
2	0.715 06	1	0.748 41	0.693 14	……	0.660 20
3	0.879 55	0.748 41	1	0.719 91	……	0.630 69
4	0.754 96	0.693 14	0.719 91	1	……	0.713 08
5	0.762 42	0.828 65	0.790 66	0.727 44	……	0.666 84
6	0.753 73	0.723 38	0.805 79	0.740 57	……	0.664 05
7	0.687 55	0.670 28	0.722 84	0.681 29	……	0.685 33
……	……	……	……	……	……	……
135	0.661 73	0.660 20	0.630 69	0.713 08	……	1

4.2 用户主题模型训练及计算

4.2.1 用户－主题模型训练 将用户看作为文档，用户所使用的标签视为文档中的词语，见表 7。利用 python 对表 7 进行 LDA 建模，取主题数 k＝5，经训练后得到用户－主题概率分布矩阵，结果见表 8。

表 7 用户－标签矩阵

用户编号	标签
A	日本；文化；社会；……；随笔
B	费孝通；乡土中国；社会学；……；日本
C	美国；日本；社会学；……；中国
D	日本；人类学；文化；……；近代史
E	社会；文化；经济；……；社会学
F	大数据；数据挖掘；互联网；……；全球化
G	历史；社会学；社会；……；失落的秘符
……	……
Y	历史；世界史；美国；……；经济学

4.2.2 相似度计算

（1）主题概率距离。在表 8 的基础上，利用公式（3）计算出两个用户在主题概率分布上的距离，结果见表 9。

表8　用户－主题概率分布矩阵

用户编号	topic1	topic2	topic3	topic4	topic5
A	0.238 71	0.002 15	0.754 84	0.002 15	0.002 15
B	0.242 67	0.296 00	0.376 00	0.002 67	0.082 67
C	0.301 49	0.092 54	0.600 00	0.002 99	0.002 99
D	0.357 01	0.076 64	0.562 62	0.001 87	0.001 87
E	0.293 20	0.079 61	0.506 80	0.001 94	0.118 45
F	0.121 57	0.003 92	0.278 43	0.356 86	0.239 22
G	0.002 47	0.125 93	0.471 60	0.397 53	0.002 47
……	……	……	……	……	……
Y	0.367 27	0.040 00	0.294 55	0.003 64	0.294 55

表9　用户间概率分布距离矩阵

用户编号	A	B	C	D	……	Y
A	0	0.153 71	0.034 01	0.035 33	……	0.166 02
B	0.153 71	0	0.065 78	0.074 14	……	0.095 61
C	0.034 01	0.065 78	0	0.001 98	……	0.128 36
D	0.035 33	0.074 14	0.001 98	0	……	0.121 48
E	0.073 44	0.040 85	0.037 37	0.039 19	……	0.039 16
F	0.262 23	0.244 01	0.267 08	0.271 38	……	0.157 26
G	0.263 96	0.253	0.233 33	0.254 74	……	0.367 2
……	……	……	……	……	……	……
Y	0.166 02	0.095 61	0.128 36	0.121 48	……	0

(2)矩阵转换。利用公式(4)将此矩阵转换为相似度矩阵，结果如表10所示：

表10　用户相似矩阵

用户编号	A	B	C	D	……	y
A	1	0.866 77	0.967 11	0.965 88	……	0.857 62
B	0.866 77	1	0.938 28	0.930 98	……	0.912 73
C	0.967 11	0.938 28	1	0.998 02	……	0.886 24
D	0.965 88	0.930 98	0.998 02	1	……	0.891 68
E	0.931 58	0.960 75	0.963 98	0.962 29	……	0.962 32
F	0.792 25	0.803 85	0.789 22	0.786 55	……	0.864 11
G	0.791 16	0.798 08	0.810 81	0.796 98	……	0.731 42
……	……	……	……	……	……	……
Y	0.857 62	0.912 73	0.886 24	0.891 68	……	1

4.3　推荐标签生成

为直观显示标签推荐过程，本文随机选取用户“enenn”并以其标注图书资源《菊与刀》为例，在上述得出资源相似度(见表6)与用户相似度(见表10)的基础上，产生基于相似资源与基于相似用户的推荐，并融合基于内容的标签推荐形成最终标签推荐结果从而阐释整个推荐标签生成过程。

4.3.1　基于相似资源的推荐　这一推荐过程的基本思想是采用基于相似资源的协同过滤，在计算出资源相似度之后对其降序排序，将与目标资源 $r \in R$ 最为相似的 m_1 个资源作为其近邻资源集合，对这些资源的热门标签加权排序，最后推荐靠前的 n_1 个标签。具体步骤如下：

(1)选择目标资源r，将其与系统中所有资源的相似度进行降序排列。

(2)选择最靠前的 m_1 个相似资源生成近邻资源集R’。

(3)将近邻资源集中所有资源的热门标签进行合并与加权排序。对于每一个标签t，其权重如公式(5)所示：

$$W(t) = \sum_{r \in R} Sim(r, r') \times Freq(t) \qquad \text{公式(5)}$$

公式(5)中，Sim(r,r’)为目标资源r与资源r’的相似度，Freq(t)为标签t的出现频率。

(4)对于排序过后的标签，选择靠前的 n_1 个作为基于相似资源的推荐结果，并将其权重归一化。

运用以上步骤，得到目标资源的近邻资源集和基于相似资源的标签推荐候选集A，如表11、表12所示：

表11　近邻资源集

相似度	《100个理由》	《武士道》	《叶隐闻书》	《中国大历史》	《乌合之众》
《菊与刀》	0.891 40	0.891 29	0.886 28	0.871 26	0.835 16

通过对目标资源《菊与刀》进行了解，发现该资源是描写日本文化以及对日本民族、历史和社会等各方面进行研究的一本书，结合表12的推荐结果可以看出，基于相似资源所推荐的标签与该资源内容的特征联系程度十分紧密，能够满足用户标注时的基本需求，且该方法未涉及到任何用户相关因素，也不会受到文本内容字数限制，因而稳定性较好，但其推荐结果比较宽泛，重复度较高，且推荐的都是系统中资源的热门标签，社会化因素高，也未考虑到用户的兴趣，无法推荐新颖的标签，可见，基于相似资源的推荐结果仍有缺陷。

表12　标签推荐候选集A

标签	文化	历史	日本	人文	日本文化	武士道	中国	日本研究	社会学	社会
权重	0.282 08	0.182 59	0.103 24	0.102 66	0.103 24	0.045 84	0.045 46	0.045 84	0.044 52	0.044 52

4.3.2 基于相似用户的推荐 在计算出用户之间的相似度后，寻找与目标用户 u∈U 最相似的且标注过目标资源的 m_2 个用户作为近邻用户集合，然后将其用于标注目标资源的标签进行加权排序，推荐靠前 n_2 个标签。具体步骤如下：

(1)选择目标用户 u，将其与系统中各用户相似度进行降序排列。

(2)选择最靠前的且标注过目标资源 r 的 m_2 个用户为近邻用户集 U'。

(3)将近邻用户集中的用户所有标注过目标资源的标签进行合并与加权排序。对于每一个标签 t，其权重如公式(6)所示：

$$W(t) = \sum_{u \in U} Sim(u,u') \times Freq(t) \qquad 公式(6)$$

公式(6)中 Sim(u,u')为目标用户 u 与用户 u'的相似度，Freq(t)为标签 t 的出现频率。

(4)对于排序过后的标签，选择靠前的 n_2 个作为基于相似用户的推荐结果，并将其权重归一化。

运用以上步骤，得到目标用户的近邻用户集和基于相似用户的标签推荐候选集 B，如表 13、表 14 所示：

表 13 近邻用户集

相似度	飞扬	路过蜻蜓	新青年	Banyan
enenn	0.998 02	0.965 88	0.950 74	0.930 98

表 14 标签推荐候选集 B

标签	日本	美国	社会学	文化	人类学	历史	人文
权重	0.484 78	0.123 78	0.121 59	0.120 80	0.118 61	0.030 44	0.029 34

分析表 14，并与表 12 和表 15 比较，该推荐方法考虑到了用户偏好的因素，因此推荐了一些个性化标签，如“美国”“人类学”这两个标签并未在基于相似资源的推荐结果(表 12)中出现，但是考虑到《菊与刀》的作者是美国的著名人类学家，该书也分析研究了日本整个民族的特性及日本人的性格特征等，因此这两个标签对于描述此书来说是十分重要的，可见基于相似用户的标签推荐能够提高标签的新颖性，而较之于基于资源内容的推荐结果(表 15)，该方法的结果精度更好，内容更全面，但也正是因为考虑到了用户因素，所以存在冷启动等问题。

4.3.3 基于内容的推荐 资源的内容特征信息可以直观地揭示资源的属性，是推荐标签的重要来源之一，且它不需要依赖用户行为信息的优点可以很好地弥补基于协同过滤技术存在的问题。本文选择 TF-IDF 特征词抽取技术作为基于内容的标签推荐方法。

TF-IDF(Term Frequency-Inverse Document Frequency)是一种在信息检索、文本分类等领域中常用的评价一个词语在文档集中对于某一文档的重要性的加权技术。TF-IDF 的基本思想是，假如一个词在某篇文档中出现频率非常高而在其他文档中出现频率极小，那么这个词对于该篇文档来说就十分重要，区分能力较好[4]。一个词的 TF-IDF 计算公式如公式(7)所示：

$$W_{ij} = tf_{ij} \times idf_i = \frac{n_{ij}}{\sum_k n_{kj}} \times \log(\frac{N}{n_i}) \qquad 公式(7)$$

公式(7)中，n_{ij}表示特征词 t_i 在文档 d_j 中出现的次数，分母则表示在文档 d_j 中所有词出现的次数总和，N 表示文档总数，n_i 表示出现特征词 t_i 的文档数[28]。可以看出，TF-IDF 值高的那些词通常是描述文档内容特征的最佳词项[29]。具体步骤如下：

(1) 利用 3.2 节中对图书简介内容进行预处理后的包含各简介特征词的文档集合，利用公式(7)计算所有词的权重；

(2) 对计算结果进行降序排序后，选择靠前的 n_3 个词作为基于资源内容关键词的推荐结果，并将其权重归一化。

经由以上步骤得到基于内容的标签推荐候选集 C，如表 15 所示：

表 15 标签推荐候选集 C

标签	日本	文化	菊与刀	本尼迪克特	矛盾
权重	0.490 81	0.187 70	0.149 02	0.099 56	0.072 92

与前两种推荐结果(表 12、表 14)相比，基于内容的标签推荐能够直接揭示资源的各种属性，如表 15 中“日本”“文化”描述了资源的内容特征，而“菊与刀”“本尼迪克特”描述了资源的书名、作者两个外部特征，因此其推荐结果精确，且不受用户因素影响，有效避免了冷启动的问题。但由于文本字数限制等原因，其推荐结果较为狭窄，不够全面，并出现了不适合描述本书内容的标签，如“矛盾”，因此，对于文本内容充足的资源来说，该方法或许会更加适用。

4.4 推荐结果生成

经过上述计算，得到的表 12、表 14、表 15 分别为目标用户 u 和目标资源 r 的基于相似资源、相似用户以及基于内容的 3 种标签推荐候选集。这 3 种结果则分别对应了标签系统中标签、用户、资源 3 种角度的标签来源，即资源热门标签、用户兴趣标签、资源内容标签。因为这三大元素共同构成了社会标注系统，所以认为它们的重要程度是相一致的，因而在对这 3 种推荐结果进行融合时，采用加权计算方法，分别将它们各

熊回香，窦燕.基于LDA主题模型的标签混合推荐研究[J].图书情报工作,2018,62(3):104－113.

自得出的结果权重系数进行归一化处理,然后将权重相加并按降序排列,选择最靠前的 n_4 标签作为最终推荐结果提交给目标用户 u。因豆瓣读书中推荐给用户的标签个数为 10 个,因此本文最终也同样向用户推荐 10 个标签。

经计算后,对于目标用户 enenn 在标注目标资源《菊与刀》时,系统最终将为其推荐的标签结果如表 16 所示:

表 16　标签推荐最终结果

标签	日本	文化	历史	社会学	菊与刀	人文	美国	人类学	日本文化	本尼迪克特
权重	1.078 83	0.590 59	0.213 03	0.166 11	0.149 02	0.132 00	0.123 78	0.118 61	0.103 24	0.099 56

4.5　结果评价与分析

为了验证推荐结果的准确度,采用精确率(Precisions)、召回率(Recall)、F_1 值为评价指标。其公式如公式(8)－(10)所示[5]:

$$Precision = \frac{TP}{TP + FP} \quad \text{公式(8)}$$

$$Recall = \frac{TP}{TP + FP} \quad \text{公式(9)}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \text{公式(10)}$$

式中 TP 表示推荐正确的标签个数,FP 表示推荐错误的标签个数,FN 表示原本应该被推荐但是却没有被推荐的个数。

随机抽取实验数据的 80% 作为训练集,剩下的 20% 作为测试集对其进行预测,统计本文实验推荐结果的准确率、召回率和 F_1 值。同时,为了进一步验证本文提出方法的有效性,计算当前几种标签推荐(基于资源内容的推荐、基于相似资源的推荐、基于相似用户的推荐)的各值并进行比较。为方便描述,将本文提出的推荐方法简称为 I + U + C,基于相似资源的推荐方法简称为 Item-CF,基于相似用户的推荐为 User-CF,基于资源内容的推荐为 Contend-Based。几种方法的比较结果如图 3 所示:

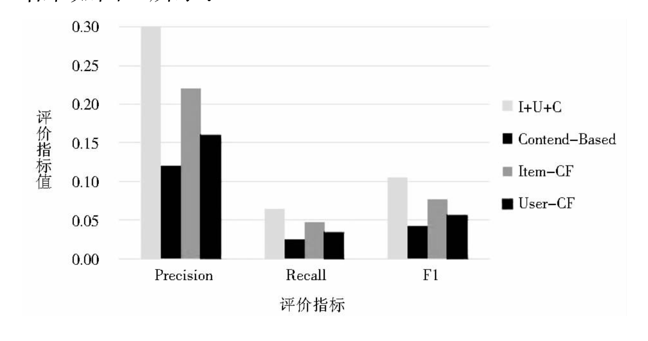


图 3　几种方法实验结果比较

表 12、表 14、表 15 分别是 3 种单一推荐方法下得到的结果,虽然各方法所推荐的标签与目标资源《菊与刀》的情况十分符合,都能明显揭示该资源特征的标签,但考虑实际标注情况,可以发现单一推荐方法都有无法避免的缺陷。

首先,从基于相似资源的推荐结果看(表 12),该方法是从标签系统中资源热门标签角度出发的,其推荐结果稳定,基本上都能很好地描述与概括资源特征,且由于其来源是热门标签,因此用户对于其可接受度高,但也正因为这样使得结果过于社会化,标签重复率高,缺乏新颖性,用户可选择的范围受限;从基于相似用户的推荐看(表 14),该结果的来源是系统中用户所使用过的兴趣标签,其考虑到了用户的兴趣偏好,结果高度准确,且有新颖性,个性化程度高,但有时却过于强调用户个性而受到影响;从基于内容的推荐看(表 15),其从资源内容标签角度出发,标签直接抽取自资源本身,不会受到社会化与用户因素影响,描述精准,更加贴近资源本身,但却会受到文本内容各方面的限制,其所推荐的结果数量是几种方法中最少的。

对比 3 种单一方法,本文提出方法的实验结果见表 16,推荐结果较为理想。其主要针对社会标注系统中,已有资源、标签的用户、信息较为充足的情况下,在用户准备标注资源时,从资源内容、资源的热门标签以及用户曾使用过的标签 3 种角度出发对相关信息进行计算,向其进行标签推荐,但不适用于信息特别分散的情况。分析表 16 的结果,该方法融合 3 种推荐技术优点并一一对应标注系统中 3 种标签来源,其所推荐的标签从多方面角度揭示了资源的特征:有来自资源热门的社会化标签(如"日本""文化"),有来自用户兴趣的个性化的标签(如"美国""人类学"),出自资源本身的关键词(如"本尼迪克特"),覆盖范围广,并且都较为规范,效果良好,同时规避了冷启动、文本限制等问题,用户在标注时可选择的标签范围也随之扩大,他们能够根据自己的需求与对资源的理解来选择合适的标签。从图 3 各方法实验结果数据对比也可以发现,本文的方法在精确率、召回率以及 F1 值这几个指标都优于其他几种方法,证明了该方法能够有效地提高推荐的准确度。

豆瓣网是目前我国社会标注网站中最为典型的代

表,不论是国内的新浪微博、知乎,还是国外的 Delicious、Flickr(图片)等网站都与豆瓣一样,其标签系统允许用户采用自定义的关键词来对资源进行标注,因此本文虽只针对豆瓣读书这一类数据集进行了实验,但只要是利用关键词作为资源的标注,本文所提出的方法都适用,有一定的推广性。

5 结语

标签推荐与一般个性化推荐不同,需要考虑用户、资源与标签 3 种因素,且现有推荐方法虽不同程度运用了协同过滤思想,但在计算相似度时却简单地以词共现为基础而忽略了标签间的语义关系,影响推荐结果的准确性。本文考虑到了标签间及资源内容间的语义信息,并从资源内容标签、资源热门标签、用户兴趣标签 3 种标签来源角度出发,融合与之对应的 3 种推荐技术,并利用 LDA 从主题语义层面计算其各自在主题概率上的分布作为相似度计算的数据基础[25],产生最后的推荐结果。从实验结果可以看出,本文提出的混合型标签推荐方法,既通过对数据的降维及运用语义关系提高了相似度计算的准确性,又能使推荐结果具有社会化及新颖性等特点,还能缓解冷启动、数据稀疏等问题,以此提高标签推荐的准确性和标签质量,一定程度上达到标签规范的目的,可以为将来的标签推荐提供一定参考。但本文为了清晰论述模型,在进行实验时所选数据量较少,部分参数使用经验值,同时也未考虑不同标签推荐个数间的差异,这必然会对实验结果有所影响。因此在后续的研究工作中,需扩大数据量,检验此方法在大规模数据集上的效果,同时改进相关算法,研究适合此模型的最佳参数值,从而优化推荐过程,提高推荐结果精确性,使理论能够更好地走向实践。

参考文献:

[1] KRESTEL R, FANKHUSER P. Tag recommendation using probabilistic topic models[C/OL]//Proceedings of ECML PKDD discovery challenge(DC09), Bled, Slovenia, 2009:131 – 141[2017 – 08 – 23]. https://www.kde.cs.uni-kassel.de/ws/dc09/papers/proceedings.pdf#page=131.

[2] 金燕,陈玉. 基于本体的标签控制方法研究[J]. 图书馆理论与实践,2010(7):26 – 29.

[3] BOGÁRDI-MÉSZÖLYÁ, RÖVID A, ISHIKAWA H, et al. Tag and topic recommendation systems[J]. Acta polytechnica hungarica, 2013, 10(10):171 – 191.

[4] 范永全,刘艳,陆园. 社会化推荐系统的研究进展综述[J]. 现代计算机:普及版,2014 (10):29 – 33.

[5] 张引. 社会标注系统中标签推荐方法研究[D]. 沈阳:东北大学,2012.

[6] 乔绿茵,张敏. 我国基于 Folksonomy 的标签推荐方法研究综述[J]. 信息资源管理学报,2012(4):41 – 46.

[7] 刘志丽. 基于内容的社会标签推荐技术研究[D]. 哈尔滨:哈尔滨工程大学,2012.

[8] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66 – 76.

[9] CAI Y, LEUNG H, LI Q, et al. Typicality-based collaborative filtering recommendation[J]. IEEE international conference on tools with artificial intelligence, 2010, 2(3):97 – 104.

[10] TATU M, SRIKANTH M, SILVA T. Tag recommendations using bookmark content[C/OL]//Proceedings of the ECML PKDD discovery challenge at 18th European conference on Machine Learning, Antwerp, Belgium, 2008:96 – 107[2017 – 08 – 23]. https://www.researchgate.net/profile/Antal_Van_Den_Bosch2/publication/228075659_Using_Language_Models_for_Spam_Detection_in_Social_Bookmarking/links/09e4150b273637375e000000/Using-Language-Models-for-Spam-Detection-in-Social-Bookmarking.pdf#page=104.

[11] MISHNE G. AutoTag: a collaborative approach to automated tag assignment for weblog posts[C]//Proceedings of the 15th international conference on World Wide Web. New York: ACM Press, 2006:953 – 954.

[12] MARINHO L, SCHRNIDTTHIEME L. Collaborative tag recommendations[C]//Data Analysis, Machine Learning -Proceedings of the 31st Annual conference of the German classification society, Albert-Ludwigs-Universität Freiburg, German, 2008:533 – 540[2017 – 08 – 23]. https://link.springer.com/chapter/10.1007%2F978-3-540-78246-9_63.

[13] HOTHO A, JASCHKE R, SCHMIZT C, et al. InformationRetrieval in folksonomies: search and ranking[J]. Lecture notes in computer science, 2006, 4011:411 – 426.

[14] 宋洪鑫. 基于标签与内容的 blog 检索实验系统研究与实现[D]. 北京:北京邮电大学,2011.

[15] 高兵. 问答式社区的标签推荐技术研究[D]. 哈尔滨:哈尔滨工业大学,2009.

[16] 王传豹. 基于协同过滤和文本相似度的标签推荐及搜索优化[D]. 保定:河北大学,2011.

[17] 安志伟. 社会标签推荐张量分解方法研究[D]. 长沙:中南大学,2011.

[18] 张亮. 基于 LDA 主题模型的标签推荐方法研究[J]. 现代情报,2016,36(2):53 – 56.

[19] 李慧宗,胡学钢,杨恒宇,等. 基于 LDA 的社会化标签综合聚类方法[J]. 情报学报,2015,34(2):146 – 155.

[20] 邸亮,杜永萍. LDA 模型在微博用户推荐中的应用[J]. 计算机工程,2014,40(5):1 – 6.

[21] yhao2014. 通俗理解 LDA 主题模型[EB/OL]. [2017 – 06 – 21]. http://blog.csdn.net/yhao2014/article/details/51098037.

[22] 张培晶,宋蕾. 基于 LDA 的微博文本主题建模方法研究述评[J]. 图书情报工作,2012,56 (24):120－126.
[23] 王振振,何明,杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学,2013,40 (12):229－232.
[24] 钟青燕,苏一丹,梁胜勇. 基于层次聚类和语义的标签推荐研究[J]. 微计算机信息,2010,26(36):199－203.
[25] 王茜,王均波. 一种改进的协同过滤推荐算法[J]. 计算机科学,2010,37(6):226－228.
[26] 熊回香. 面向 Web3.0 的大众分类研究[D]. 武汉:华中师范大学,2011.
[27] 豆瓣读书[EB/OL]. [2017－06－05]. https://book.douban.com/.
[28] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29(6):167－170.
[29] ANAND R,JEFFREY D. 大数据·互联网大规模数据挖掘与分布式处理[M]. 北京:人民邮电出版社,2012:6－7.

作者贡献说明:
熊回香:提出研究方向,确定研究方法和论文的逻辑架构;
窦燕:获取数据和撰写论文。

Research on Tag Hybrid Recommendation Based on LDA Topic Model

Xiong Huixiang　Dou Yan
School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] For the current tag recommendation methods' results not satisfied, this paper aims to improve the traditional similarity calculation method and combine a variety of tag recommendation methods to improve the recommended accuracy. [**Method/process**] Based on the idea of content and collaborative filtering, LDA is used to calculate the similarity then find the neighbor of resources and users, and combine keywords which are extracted from resource contents to construct the tag hybrid recommendation model. Finally, "Douban reading" is taken as an example to verify the model's effectiveness and compared with several tag recommendation methods. [**Result/conclusion**] In the social tagging system, three dimensions including user, resource and tag should be considered. Only from one single angle will inevitably cause incomplete results . At the same time, the introduction of LDA in similarity calculation can exploit the potential semantic relation and improve the recommended quality. And the combination of a variety of ways to learn from each other can make the results more satisfactory.

Keywords: social tagging　tag recommendation　collaborative filtering　LDA